Analysis of high-dimensional data with sparse structure

Author(s):
Meinshausen, Nicolai

Publication date:
2005

Permanent link:
https://doi.org/10.3929/ethz-a-005081993

Diss. ETH No. 16244

Analysis of High-Dimensional Data with Sparse Structure

A dissertation submitted to the
SWISS FEDERAL INSTITUTE OF TECHNOLOGY
ZURICH

for the degree of
Doctor of Mathematics

presented by
NICOLAI FELIX MEINSHAUSEN
Dipl. Phys., ETH Zürich
MSc, University of Oxford
born May 9, 1976
citizen of Germany

accepted on the recommendation of
Prof. Dr. Peter Bühlmann, examiner
Prof. Dr. Hans Rudolf Künsch, co-examiner

2005

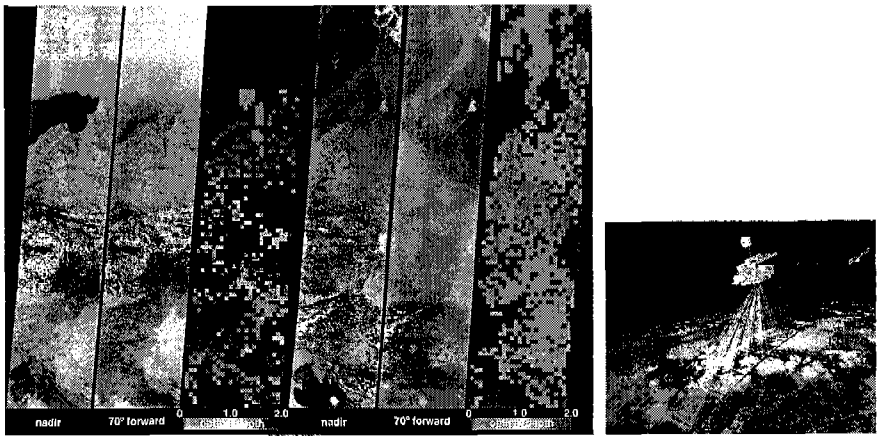

Figure 1: *Aerosol measurements over Central and Eastern Europe (left) with the MISR-instrument. The MISR-instrument is part of the TERRA satellite (right), orbiting earth in a sun-synchronous orbit. Measurements are made for four different bands in nine angles, providing data at a rate of 3.3 Megabits/second. The goal of the mission is to gain a deeper understanding of the vital contribution of aerosols and clouds to global climate dynamics. Pictures and further information from http://www-misr.jpl.nasa.gov. The high-dimensionality of the data is evident, possible sparse structures maybe less so.*

Abstract

This thesis is a collection of papers and manuscripts about analysis of high-dimensional data with sparse structure. The focus of this thesis is on prediction and multiple testing.

The terms "sparse structure" and "high-dimensional" can have quite a broad range of connotations. Instead of generally valid definitions, I will point out some relevant applications to clarify the spirit in which these terms are used in this thesis.

The dimensionality of a prediction problem is usually defined as the number p of available predictor variables. Setting p into relation with the number n of observations which are available for estimation or training, a problem is said to be "high-dimensional" if the number of predictor variables p is much larger than the number n of available observations. Consider as examples land-use classification, cloud detection, or aerosol concentration estimation with satellite-based measurements, as briefly explained in the caption of Figure 1, or Microarray gene expression data, Figure 2. The sample size for the latter type of experiment is typically in

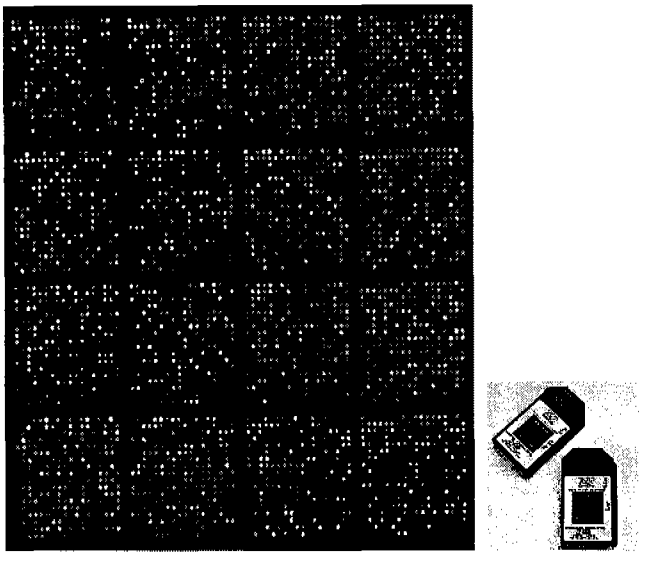

Figure 2: *Microarray showing patterns of gene expression influenced by nicotine (left), from Powledge (2004). Gene Chips (right) are one possibility for gene expression measurement.*

the order of dozens or hundreds, while the number of genes on the chip is in the order of thousands or tens of thousands. For cloud detection using satellite data, a lot of information from different spectral bands and viewing angles is available, while the number of (hand labelled) training data for cloud classification is usually very small, as these training data can only be obtained by manual classification, pixel by pixel.

We examine the asymptotic properties of suitable prediction methods in a setting that reflects the "large p, small n" situation. The number of variables $p = p_n$ grows with n in the asymptotic analysis, possibly very fast, so that $p_n \gg n$ for $n \to \infty$.

Crucially, one has to assume in this setting that the data have "sparse structure" in some sense, meaning that most of the predictor variables are irrelevant for accurate prediction. The task is hence to filter out the relevant subset of predictor variables. While high dimensionality of a dataset is evident from the start, it is usually not easy to verify structural sparseness. Often, sparseness is an assumption one has to make in the high-dimensional case, as it is almost impossible to analyze non-sparse high-dimensional data. In the words of Friedman et al. (2004), this is termed the "bet on sparsity":

> *Use a method that does well in sparse problems, since no procedure does well in dense problems.*

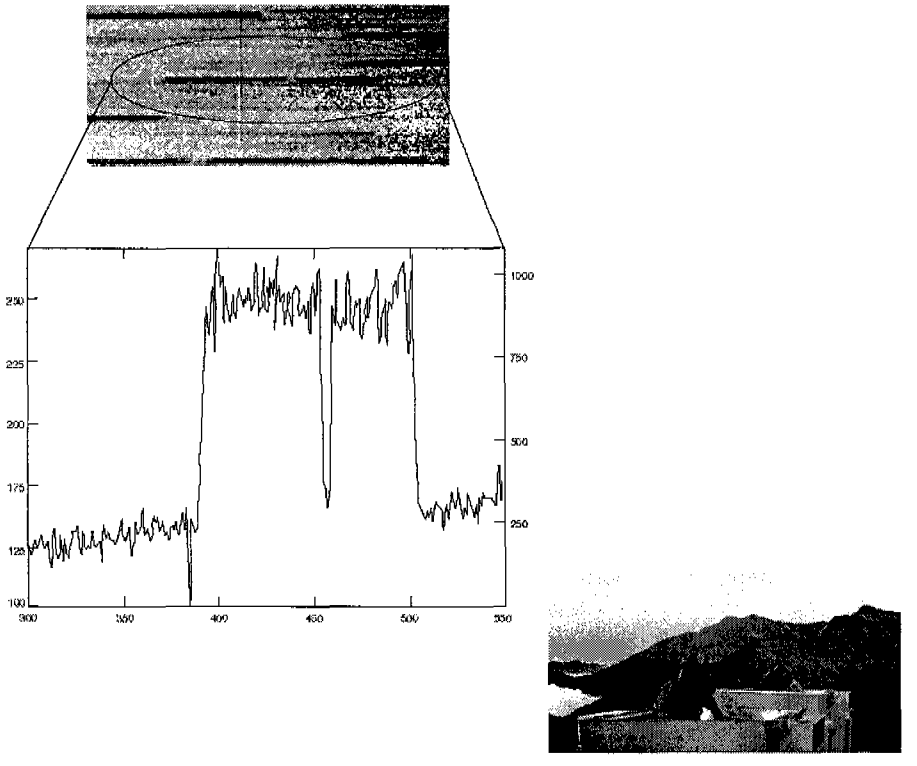

Figure 3: *Telescopes in Taiwan (right), tracking 2-3000 stars in clear nights to detect occultations by Kuiper Belt objects as part of the Taiwanese-American Occultation Survey (TAOS). Over the course of a year,* $10^{11} - 10^{12}$ *signals have to be tested for the presence of occultations, which are expected to total in between a few dozen to a few thousand among all these measurements. Example trailed image data (left), where the top panel shows a portion of CCD exposure; trailed stars are clearly seen, while the bottom panel shows the raw photometric light curve of the indicated star. Source of the pictures and further information under http://taos.asiaa.sinica.edu.tw. Detection of weak signals among a large number of noisy measurements is discussed in Chapter 4, with application to data from the TAOS project.*

In Chapter 1, the asymptotic properties of Lasso-estimates are examined. In particular, it is shown that the Lasso is consistent for variable selection under an appropriate choice of the penalty parameter. This chapter led to the paper Meinshausen and Bühlmann (2006). However, it is shown as well that the prediction-optimal choice of the penalty parameter selects too many variables and is not consistent for variable selection. Moreover, the rates of convergence of the l_2-loss can be slow for high-dimensional data with the Lasso, as shown in Chapter 3. An improvement of Lasso estimation is proposed, which leads to faster convergence rates for high-dimensional problems. Moreover, the prediction-optimal choice of the penalty parameters leads now to consistent variable selection. Chapter 3 corresponds to manuscript Meinshausen (2005a).

In Chapter 1, it was shown that Graphical Gaussian Models can be estimated consistently by sequentially applying the Lasso to all nodes in the graph and combining neighborhood estimates into one estimate of the graph. However, neighborhood estimates are not necessarily consistent with each other. Even though asymptotically this poses no problem, it has a certain appeal trying to maximize directly the negative log-likelihood of the Gaussian distribution under a l_1-penalisation of elements of the inverse covariance matrix. However, it is shown in Chapter 2 that this procedure is not consistent even if the number of nodes in the graph is constant for an increasing number of observations (Meinshausen, 2005b).

The second part of the thesis is about multiple testing. The terms "high-dimensionality" and "sparse structure" have slightly different meaning in this context. The dimensionality of the data corresponds to the number of hypotheses, while "sparse structure" refers to just very few hypotheses being false null hypotheses (very few signals amidst a lot of noise). Detection of such rare and possibly faint events is rather challenging. Sparsity is thus a challenge for multiple testing, while it is a necessity for accurate prediction.

In this part, the two main applications are a signal detection problem in astronomy, as shown in Figure 3, and, again, analysis of microarray gene expression data. The basic question that is discussed for these multiple testing problems is how to estimate the proportion of false null hypotheses among a large number of tests. For the astronomical application, the number of false null hypotheses translates directly into an estimator for the number of objects of a certain size in the Kuiper Belt. Interestingly, it turns out that it is possible to estimate a lower bound for the proportion of false null hypotheses without being able to tell which of them are false null hypotheses.

An estimator for independent tests is proposed in Chapter 4. This estimator is shown to improve on existing techniques and optimality of the procedure is proven. The results of Chapter 4 are published in (Meinshausen and Rice, 2006). Chapter 5 treats the case of dependent test statistics. Using a permutation-based approach, the results of Chapter 4 can be generalized. The results of this chapter are published in (Meinshausen and Bühlmann, 2005). Finally, it is shown that the procedures of Chapters 4 and 5 can not only be used to control the total number of false null hypotheses, but also to control the proportion of false discoveries, simultaneously for all rejection regions (Meinshausen, 2004).

ZUSAMMENFASSUNG

Diese Doktorarbeit hat die Analyse hoch-dimensionaler Daten zum Thema, unter Ausnutzung dünn besetzter Strukturen. Die Bedeutung der Begriffe "hoch-dimensional" und "dünn besetzt" ist im Allgemeinen nicht sehr klar umrissen. Anhand von einigen Beispielen will ich aufzeigen, wie diese Begriffe in dieser Arbeit aufgefasst und verwendet werden.

Die Dimensionalität von Daten, die für die Vorhersage einer skalaren Zielgrösse verwendet werden, ist üblicherweise definiert durch die Anzahl erklärender Variablen p. Ein Problem wird als "hoch-dimensional" bezeichnet, wenn die Anzahl p der erklärenden Variablen die Anzahl n unabhängiger Beobachtungen deutlich übersteigt.

Als ein Beispiel sei hier die Klassifikation von Bodennutzung, Wolkendetektion, oder Aerosol-konzentrations Schätzungen mithilfe von Satellitendaten genannt (Figure 1); ein weiteres Beispiel ist die Analyse von Microarray-Daten (Figure 2). Bei Microarray Daten liegen typischweise nur ein paar Dutzend oder Hunderte unabhängiger Beobachtungen vor, während die Anzahl an Genen auf dem Chip in die Tausende oder Zehntausende geht.

Es werden die Eigenschaften geeigneter Vorhersageverfahren in einer asymtotischen Weise untersucht, die die Eigenschaft "grosses p, kleines n" angemessen berücksichtigt. Die Anzahl an erklärenden Variablen $p = p_n$ wächst in dieser Analyse mit der Anzahl Beobachtungen n (möglicherweise sehr schnell), sodass $p_n \gg n$ für $n \to \infty$.

In solch einer asymptotischen Analyse muss angenommen werden, dass die Daten "dünn besetzt" sind, in dem Sinne dass viele erklärende Variablen irrelevant sind für die Vorhersage der Zielgrösse. Eine Aufgabe besteht also darin, die relevanten Variablen aus der gesamten Menge der Variablen herauszufiltern.

Eine hohe Dimensionalität der Daten ist sofort ersichtlich; eine dünne Besetzung der Daten hingegen nicht. Meistens muss diese dünne Besetzung einfach angenommen werden, da es nahezu unmöglich ist hochdimensionale Daten zu analysieren, die nicht dünn besetzt sind. Dies wird als "bet on sparsity" in Friedman et al. (2004) bezeichnet: "Use a method that does well in sparse problems, since no procedure does well in dense problems."

Die asymptotischen Eigenschaften der Lasso Schätzer werden in Kapitel 1 untersucht. Es wird gezeigt, dass Lasso konsistent für die Variablenselektion ist unter der Voraussetzung, dass der Bestrafungsparameter richtig gewählt ist. Diese Resultate sind in Meinshausen and Bühlmann (2006) veröffentlicht. Es wird zudem gezeigt, dass Lasso nicht konsistent

ist für die Variablen-selektion, wenn der Bestrafungsterm auf eine optimale Vorhersage ausgerichtet ist. In diesem Fall werden zuviele Variablen selektiert. Ausserdem kann die Konvergenzrate des l_2-Verlustes sehr langsam sein für hoch-dimensionale Daten. Dies wird in Kapitel 3 gezeigt. Eine Verbesserung des Lasso Schätzers wird vorgeschlagen, die zu schnelleren Konvergenzraten für hoch-dimensionale Daten führt. Der für die Vorhersage optimale Bestrafungsparameter führt jetzt auch zu einer konsistenten Variablen-selektion (Meinshausen, 2005a).

Es wurde gezeigt, dass Gauss'sche Graphen konsistent geschätzt werden können, indem Lasso sequentiell auf alle Knoten des Graphen angewendet wird und die Schätzungen der Nachbarschaften zu einer Schätzungen des gesamten Graphen kombiniert werden. Schätzungen der Nachbarschaften sind jedoch nicht notwendigerweise miteinander verträglich. Dies stellt asymptotisch kein Problem dar. Es erscheint attraktiv, den Graphen direkt zu schätzen durch eine Maximierung der Likelihood unter einer l_1-Bestrafung der Elemente der inversen Kovarianzmatrix. Es wird in Kapitel 2 gezeigt, dass dieses Verfahren nicht konsistent ist, selbst wenn die Anzahl Knoten des Graphen konstant ist für eine zunehmende Anzahl von Beobachtungen (Meinshausen, 2005b).

Der zweite Teil der Doktorarbeit handelt vom multiplen Testen. Die Begriffe "hoch-dimensional" und "dünn besetzt" haben eine leicht andere Bedeutung in diesem Kontext. Mit der Dimensionalität der Daten ist in diesem Teil die Anzahl an Hypothesen (oder Tests) gemeint, während "dünn besetzt" bedeutet, dass viele dieser Hypothesen die Nullhypothese erfüllen. Die Detektion von seltenen und schwachen Signalen ist eher schwierig. Somit sind dünn besetzte Daten eine Herausforderung im Bereich des multiplen Testens, jedoch eine Notwendigkeit und erwünscht für Vorhersage-probleme.

Zwei Anwendungen haben als Motivation für die Untersuchungen in diesem Teil der Arbeit gedient: zum einen ein Signal-Detektions Problem in der Astronomie (Figure 3) und zum anderen die Analyse von Genexpressions Daten.

Die grundlegende Fragestellung für beide Anwendungen ist die Schätzung der Anzahl falscher Nullhypothesen unter einer grossen Menge an durchgeführten Tests. Interessanterweise ist es möglich, die Anzahl falscher Nullhypothesen nach unten sehr präzise abzuschätzen, ohne aussagen zu können, welche Nullhypothesen falsch sind.

Ein Schätzer für unabhängige Teststatistiken wird in Kapitel 4 vorgeschlagen. Es wird gezeigt, dass dieser Schätzer bestehende Techniken verbessert und die Optimalität wird nachgewiesen. Dieses Kapitel ist in

(Meinshausen and Rice, 2006) veröffentlicht. Abhängige Teststatistiken werden in Kapitel 5 untersucht. Mithilfe eines Permutations-Ansatzes können die Resultate von Kapitel 4 verallgemeinert werden (Meinshausen and Bühlmann, 2005). Zuletzt wird gezeigt, dass die Prozeduren von den Kapiteln 4 und 5 ebenfalls dazu benutzt werden können, die Anzahl falscher Nullhypothesen für alle Verwerfungsbereiche gleichzeitig zu kontrollieren (Meinshausen, 2004).